图书情报工作
第 66 卷 第 12 期　2022 年 6 月

科学数据集知识扩散特征探析*
——以基因表达数据集为例

杨宁[1,2]　张志强[1,2]
[1] 中国科学院成都文献情报中心　成都 610041
[2] 中国科学院大学经济与管理学院图书情报与档案管理系　北京 100190

摘　要：[目的/意义]通过研究科学数据集的知识扩散特征和规律，探究其在学科发展过程中的实际作用，为科学数据集科技评价及管理政策制定提供参考。[方法/过程]以 GEO 数据库的数据集和 PubMed Central 数据库中重用数据集的全文数据为分析对象，采用内容分析法结合扩散广度、扩散强度、扩散速度等知识扩散指标对科学数据集的知识扩散特征进行探析。[结果/结论]研究结果表明，科学数据集的知识扩散广度和知识扩散强度日益加大，重用数据可以加快知识扩散速度，我国在全球科学数据领域的地位不断提高。

关键词：科学数据集　知识扩散　特征探析　测度指标
分类号：G250
DOI：10.13266/j.issn.0252-3116.2022.12.008

1　引言

科学数据是科研活动过程中产生或再加工得到的数据资料，主要类型包括实验数据、观测数据和统计数据等。其中，任意单位的数据都可以称为科学数据，而为了特定研究目的创建、收集和整理的相关科学数据集合或产品则构成科学数据集。随着数据驱动研究范式在各个学科领域的广泛普及，科学数据集逐渐成为贯穿科研过程的重要研究对象和产出结果，这些通过实验或观测得到的数据资料不但加快了科研进程，其含有的知识价值也在数据集共享和重用的过程中得到了更广泛的传播、继承和创新，实现了知识扩散。知识扩散指知识通过一定的载体进行的跨时空流动过程，通过这种知识的吸收和重组，促进了新知识的产生和科学的创新发展[1]。探索分析科学数据集的知识扩散情况，对拓展知识扩散研究范围、深层次了解科学数据集的学术价值、推进数据引用规范化、促进数据共享和重用等都具有极其重要的现实意义。

1924 年，卡耐基基金会的 W. S. Learned 在《美国公共图书馆与知识扩散》一书中首次对知识扩散进行了研究[2]。目前，国内外学者围绕知识扩散开展的研究主要可以分为三类：①知识扩散单元研究，基于各类知识扩散单元（论文、专利、作者、期刊、学科等）开展的知识扩散特征和规律的研究。黄鲁成等[3]提出了一个基于专利全引用信息的技术知识扩散特征研究框架，通过利用专利引用关系，可以从技术知识的利用和传播两个角度探索技术知识扩散特征。赵蓉英等[4]通过构建作者知识扩散网络，发现作者知识扩散过程与规律，并对作者知识扩散贡献程度进行了评定。岳增慧等[5]以社会网络学科为研究对象，探讨了学科知识扩散的特征。王静静等[6]通过探析国际数字人文研究中的跨学科知识扩散趋势，发现图书情报等初始相关学科的核心度正在下降，而艺术与人文、工程学等在学科研究中的地位渐趋重要。②知识扩散指标研究，通过计量或网络指标开展知识扩散的测度研究。Y. X. Liu 等[7]基于 ESI 学科分类，提出了学科知识扩散广度、强度和速度等指标。俞立平等[8]参照 h 指数的计算方法提出了用于反映学术期刊知识扩散深度的 CJH 指标。H. Nakamura 等[9]提出利用施引文献发表时间与被引文献发表时间之间的差值作为知识扩散延时指标。宋歌[10]利用扩散理论、社会网络分析和引文分析方法，从知识网络结构特征角度提出了创新扩散广度、

* 本文系国家社会科学基金重点项目“面向领域知识发现的学科信息学理论与应用研究”（项目编号：17ATQ008）研究成果之一。
作者简介：杨宁，副研究馆员，博士研究生；张志强，研究员，博士生导师，通信作者，E-mail：zhangzq@clas.ac.cn。
收稿日期：2021-12-16　修回日期：2022-03-12　本文起止页码：82-91　本文责任编辑：杜杏叶

杨宁，张志强. 科学数据集知识扩散特征探析——以基因表达数据集为例[J]. 图书情报工作,2022,66(12):82 –91.

速度、强度及延时等测度指标。③知识扩散模型研究，通过各类模型开展知识的扩散和演化过程研究。I. Z. Kiss 等[11]参考传染病模型提出一种基于个体的有向加权知识扩散模型,该模型可以用于描述研究主题在不同学科之间的扩散过程。X. Gao 等[12]综合网络分析、引文分析和可视化的方法,提出一种基于引文的时序网络知识扩散模型,该模型综合了社会网络分析、网络可视化、引文分析的方法,可以从网络结构视角揭示知识扩散的过程。

从上述研究可以看出,当前知识扩散研究大多以论文、专利、作者等载体为主,以引用、共现等关系构建网络,通过计量或网络分析方法来研究知识扩散的特征和规律。虽然近几年出现了以图书[13]、软件[14]、基金[15]等为知识单元的知识扩散研究,但还少有研究围绕科学数据集这一科研成果进行知识扩散特征分析,究其原因主要有两点:一是当前缺乏统一的数据引用标准规范,科学数据集在论文中常以提及等非规范引用形式出现,科学数据集的引用情况难以追溯和统计[16]。研究发现数据引文索引(Data Citation Index, DCI)收录的科学数据集存在很大比例的零被引现象,数据集知识扩散广度和深度也十分有限[17],这些问题给基于引用关系的科学数据集相关研究造成了极大的误差和困难;二是当前科学数据集引用的研究仍然以抽样调查和人工内容分析方法为主[18 –20],研究的文献数量和范围十分有限,研究层次也未深入到科学数据集的元数据信息,难以归纳总结出宏观层面的特征规律,无法应用于知识扩散方面的研究。

本研究将以生物医学领域的基因表达数据集作为研究对象,利用数据集的元数据获取作者、机构及数据集公开日等信息,运用内容分析法识别重用数据集的文献信息,建立科学数据集和文献间的引证关系,并利用知识扩散相关测度指标探析科学数据集在学术交流体系中的扩散特征。本研究的意义在于:一方面,将科学数据引入知识扩散研究领域,拓展并丰富知识扩散理论及方法研究,深化对科学数据知识扩散过程的认识和理解;另一方面,丰富并扩充了科研评价的内容和应用领域,为科学数据管理和服务提供全新视角的参考,也可为后续研究提供新的思路。

2 研究方法

2.1 基本思路

本研究涉及的科学数据集数据来自基因表达综合数据库(Gene Expression Omnibus, GEO),该数据库是由美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)创建并维护的全球性高通量分子丰度数据库[21],同时也是当前全球存储规模最大、数据最全面的基因表达数据库,收录了世界各国研究者提交并共享的基因芯片数据和高通量测序数据。GEO 数据库将用户或科研人员递呈和共享的数据进行分类存储并为其分配一个唯一且永恒不变的登录号(Accession Number),并要求共享数据的研究文献在公开发表后,将数据进行公开便于其他科研人员利用该数据进行后续研究。本研究将首先获取 GEO 数据库数据集的元数据信息,对数据集进行多视角的计量分析和变化趋势分析。

科学数据集知识扩散特征研究主要基于数据重用关系,数据重用也被称为数据复用或数据二次分析,相关研究开始于 20 世纪 90 年代,主要指为了重现研究结果或新的研究目的而将以往原始数据集或再组合的数据集进行再分析的过程[22]。本研究通过 PubMed Central(PMC)获取生物医学领域的科学文献全文数据,PMC 是 NCBI 提供的免费生物医学期刊文献全文数据库[23]。本文将利用规则抽取登录号的方式,在全文中识别并获取数据集的使用信息,并将发表时间晚于数据集公开时间的文献定义为重用数据集的文献,从而获得重用数据集的文献信息。最终利用数据集、重用数据集的文献及二者间建立的引证关系,进行科学数据集知识扩散特征的分析和研究。整体研究思路如图 1 所示:

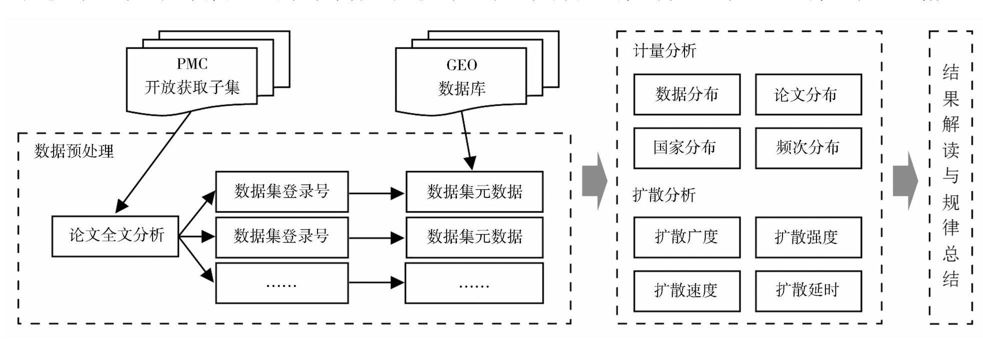


图 1 整体研究思路

2.2 数据获取

GEO 的原始数据分为平台(platform)、样本(sample)和系列(series),它们被分别保存在三种独立但具备关联性的实体数据库中。其中,平台包含芯片或测序平台的描述信息和注释信息,通常包含多个提交者提交的样本;样本用于记录单个样本的基因表达测量数据信息,是原始实验结果的基本单位;系列是由多个样本组成的具有生物学意义的数据集。此外,GEO 数据库根据原始数据的观测角度,分别从"实验"和"基因"的角度将原始数据分类整理并放置在数据集(Datasets)和表达图谱(Profiles)两个数据库中,本文通过 GEO 数据库检索并获取了以上 3 类原始数据的全部数据集信息。

此外,本文通过 PMC 提供的 FTP 服务批量下载了 2021 年 5 月 25 日前的文件包,将索引文件合并后获取到文献的基本信息及本地文件位置,并利用 Python 对 PMC 全文数据进行了解析处理,最终,共获取到 3 219 908篇全文文献。科学数据集的使用识别采用基于模式匹配的方法,通过正则表达式在全文文本中进行抽取。3 种原始数据的数量、主要元数据信息、抽取正则表达式规则等科学数据集的基本信息如表 1 所示:

表 1　GEO 科学数据集基本信息

数据集	元数据	数量/个	抽取规则
平台	公开日、标题、技术类型、物种、联系人、国家…	23 965	GPL\d +
样本	公开日、标题、样本类型、平台、联系人、国家…	4 716 270	GSM\d +
系列	公开日、标题、物种、贡献者、原文信息、联系人、国家…	158 368	GSE\d +

对抽取结果进行分析后发现,部分文献还存在如"GSE4357 – GSE4380"或"GSE4357 to GSE4380"等形式的数据集批量使用行为,需要单独构建批量抽取规则,并设置最大抽取阈值为 500,超出阈值范围的不进行抽取,从而提取出批量使用的数据集登录号。本研究将发表日期在数据集公开日期之后的文献定义为重用文献,经过识别抽取后发现,共有 39 189 篇文献重用了 GEO 数据库中数据集,数据集总量为 57 841 个,重用频次合计 294 517 次,存在 GEO 数据集重用行为的文献数量占全部文献数量的 1.22% 。

2.3 科学数据集知识扩散测度指标

结合以往研究在其他知识扩散单元的指标定义,结合科学数据集自身的特点,本研究提出了数据知识扩散广度、数据知识扩散强度、数据知识扩散速度及数据知识扩散延时 4 个测度指标:

(1)数据知识扩散广度(Data Knowledge Diffusion Breadth,DKDB),该指标从覆盖范围视角对数据知识扩散情况进行分析,即重用数据集的论文数量越多,则数据知识扩散广度越大,数据集的知识接收者越多。2002 年,I. Rowlands[24] 最早提出了知识扩散广度的测度指标。随后,T. F. Frandsen[25]、邱均平[26] 等对知识扩散广度等测度指标进行了修正和扩展。本文参考前人研究提出数据知识扩散广度指标 *DKDB*,其计算公式如公式(1)所示:

$$DKDB = N_i / Y_{pub} \qquad \text{公式(1)}$$

其中,N_i 表示统计年度中重用该年公开数据集的论文数,Y_{pub} 表示数据集年龄。由于知识扩散是一个动态累积过程,本文还对数据累积知识扩散广度($DKDB^*$)进行了考察,其计算公式如公式(2)所示:

$$DKDB^* = \frac{\sum_l^i N_i}{\sum_l^n N_i} \qquad \text{公式(2)}$$

其中,$1 \leqslant i \leqslant n$,$N_i$ 表示重用第 i 年公开数据集的论文数,n 为总统计年数。利用这两个指标可以反映出数据知识扩散广度在不同年份以及逐年累积的发展变化趋势。

(2)数据知识扩散强度(Data Knowledge Diffusion Intensity,DKDI),该指标从重用频次视角对数据知识扩散情况进行分析,即重用数据集的次数越多,则数据知识扩散强度越大,数据集对知识接收者的影响越大。与数据扩散广度测度方法类似,本文将从数据知识扩散强度指标 DKDI 和数据累积知识扩散强度 $DKDI^*$ 两个角度,对数据扩散强度情况进行考察和分析。其计算方法如公式(3)和公式(4)所示:

$$DKDI = N_j / Y_{pub} \qquad \text{公式(3)}$$

其中,N_j 表示统计年度中该年公开数据集被论文重用的总频次,Y_{pub} 表示数据集年龄。同样,本文也对数据累积知识扩散强度($DKDI^*$)进行了考察,其计算方法如公式(4)所示:

$$DKDI^* = \frac{\sum_l^j N_j}{\sum_l^n N_j} \qquad \text{公式(4)}$$

其中,$1 \leqslant j \leqslant n$,$N_j$ 表示重用第 j 年公开数据集的总频次,n 为总统计年数。利用这两个指标可以反映出数据知识扩散强度在不同年份以及逐年累积的发展变化趋势。

(3)数据知识扩散速度(Data Knowledge Diffusion Speed,DKDS),该指标从单位时间里传播距离的角度对知识扩散情况进行分析。2005 年,R. Rousseau 提出"平均扩散速度"指标,指的是一篇论文发表后,引用

该论文的期刊数量与论文年龄的比值，此后该指标也一直在被扩展和完善[27]。本文参考该指标，提出数据知识扩散速度指标，指的是一个数据集公开后，刊载了重用该数据集论文的期刊数量与数据集年龄的比值，其计算方法如公式(5)所示：

$$DKDS_y = \frac{(P_i / Y_{pub})}{m} \qquad 公式(5)$$

其中，$DKDS_y$ 为某一年公开数据集的平均数据知识扩散速度，P_i 表示所有刊载了重用该年公开数据集论文的期刊数量，Y_{pub} 表示数据集年龄，m 为该年度公开的数据集数量。

(4)数据知识扩散延时(Data Knowledge Diffusion Delay, DKDD)，该指标参考反映论文知识扩散速度的"引文滞后"指标[28]，即数据集被第一次重用的时间与数据集公开时间的时间差，从效率的角度揭示数据集的扩散速度，其计算方法如公式(6)所示：

$$DKDD = T_1 - T_0 \qquad 公式(6)$$

其中，T_1 表示第一次重用该数据集的论文发表年份，T_0 表示该数据集的公开年份。

3 结果与分析

3.1 科学数据集基本信息

GEO 数据库中被重用的数据集公开时间分布于 2000－2021 年，期间共有 57 841 个数据得到重用，总被重用频次 294 517 次，平均被重用频次 5.092 次。论文方面，存在 GEO 数据集重用行为的文献发表时间分布于 2004－2021 年，论文数量合计为 39 189 篇，平均每篇文章重用数据集 1.476 个，这些论文发表在 1 337 本期刊上。其中，最早的重用记录可以追溯到 2004 年 M. V. Osier 等的研究[29]，该研究使用了 GEO 数据的 4 个数据集 GPL205、GPL218、GPL229 和 GPL356，用于测试其提出的微阵列数据分析方案的可行性。具体的年度分布及重用情况如表 2 及图 2 所示：

表 2 GEO 数据集公开年度分布及重用情况

年度	数据集	论文数	频次	期刊数	年度	数据集	论文数	频次	期刊数
2000	12	70	512	6	2011	5 269	3 338	29 050	362
2001	84	166	1 251	26	2012	5 988	3 029	26 345	402
2002	203	673	3 548	52	2013	5 976	3 244	28 022	389
2003	286	1 375	4 721	54	2014	4 698	3 100	23 838	413
2004	723	539	3 314	107	2015	4 735	3 011	23 209	417
2005	1 272	1 034	8 088	137	2016	4 506	2 453	17 418	414
2006	1 604	1 367	10 952	178	2017	4 581	2 452	16 752	405
2007	2 230	1 674	14 452	216	2018	3 073	1 913	11 252	382
2008	2 412	2 056	16 423	254	2019	2 452	1 486	7 322	352
2009	2 686	2 242	19 346	270	2020	1 261	850	2 822	225
2010	3 687	3 039	25 734	323	2021	103	78	146	53

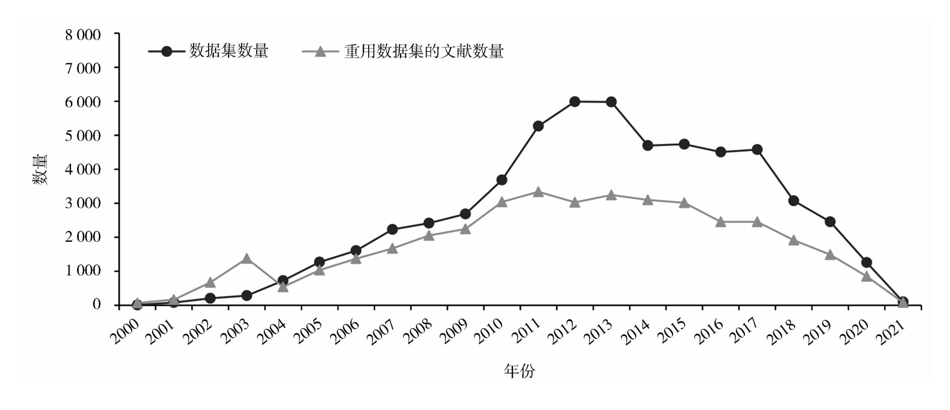


图 2 数据集及重用数据集的文献数量年度分布

由于生物医学领域从实验到论文发表存在一定滞后性，因此目前被重用较多的数据集主要发布于 2008－2017 年间，这与科研范式的转变及学科信息学等数据驱动型学科的兴起密切相关。对数据集的重用论文篇数分别统计并排序后发现，重用篇数为 1 的数据集为 43 217 个，占总数的 74.72%。而被重用最多的数

据集是出自美国著名生物芯片公司 Affymetrix 的商业数据集 GPL570,共有 1 634 篇论文重用了该数据集。如以数据集被重用次数为 X 轴,数据集数量为 Y 轴,可以得到图 3 的二者关系图。从图 3 可以较为明显地看出,数据集重用次数与数据集数量之间满足幂律分布(R^2 =0.99),绝大部分数据集只得到了少量重用,而少数数据集则得到了大量重用。

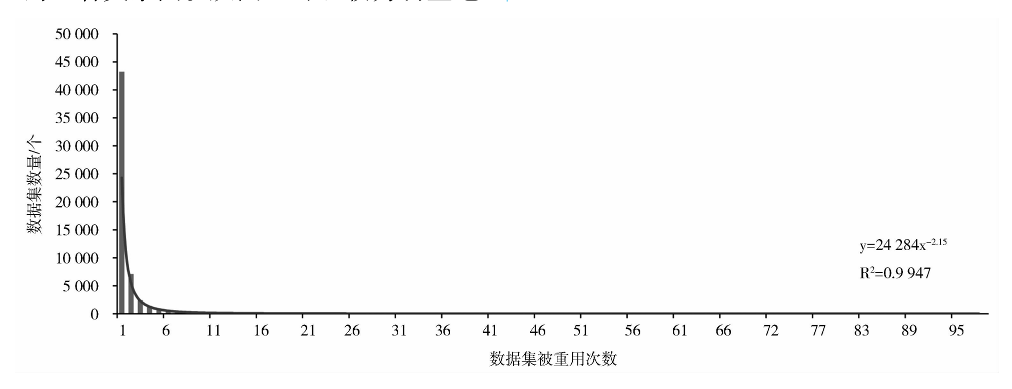


图 3　数据集数量与其被重用次数的关系

从国家和地区角度来看,共有 53 419 个数据集标注了贡献者的国籍,其中美国以 27 187 个数据集的发布量排名第一,中国以 3 976 个数据集的发布量排名第二,其他发布数据集较多的国家还包括德国、日本、英国和澳大利亚等。数据发布量在前十位国家的年度发布数量占比变化如图 4 所示:

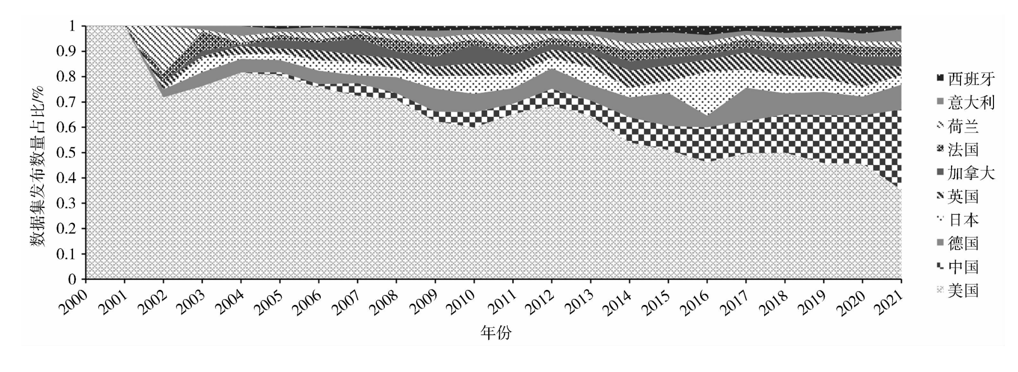


图 4　数据发布量排名前 10 国家的年度发布数量占比变化

在这些数据集中,较早出现的是美国在 2000 年和 2001 年发布的 16 个数据集。在 2002 年以后,发布数据集并获得重用的国家逐渐增多。而我国最早发布并获得重用的数据集出现在 2005 年,是由香港中文大学发布的香菇基因表达谱数据[30]。随着我国科学数据共享政策的不断出台,科研人员提升数据共享的意识也在不断提升,我国共享的数据占比在不断加大,从 2005 年只占 1.35% 到近几年已经超过 1/3 的比例。

3.2　数据知识扩散广度

数据知识扩散广度从扩散范围角度对知识扩散情况进行考察,即重用数据集的论文年度分布及变化趋势。对 GEO 数据集的数据知识扩散广度进行测算,结果如表 3 所示。2000 年数据知识扩散广度仅为3.182,而后逐年增加,2019 年达到峰值 495.333,年平均扩散广度为 233.534。

本文将计算出的 *DKDB* 取 1 000 为底的对数作为当年数据知识扩散指标,同时对当年数据知识扩散与数据累积知识扩散指标的结果进行比较和分析,见图 5。可以看出,在 2000 – 2019 年间的当年数据知识扩散广度保持着波动增长的态势,其后开始有所回落。数据累计知识扩散广度呈现 S 曲线形态,说明科学数据集逐渐在学科内产生影响力并持续受到关注,推动着生物医学领域的知识融合与创新发展,但由于数据集公开和重用具有一定的滞后性,近几年公开的数据集的数据知识扩散广度增长趋势有所减缓。

3.3　数据知识扩散强度

传统知识扩散强度主要考察的是某学科的知识单元对其他学科的影响程度,由于本文研究的数据集和重用论文都集中于生物医学领域,因此本文提出的数据知识扩散强度主要从重用频次角度对知识扩散情况进行考察,即重用数据集次数的年度分布及变化趋势。

表 3　GEO 数据集的数据知识扩散广度

年度	N_i	DKDB	DKDB*	$Log_{1000}DKDB$	年度	N_i	DKDB	DKDB*	$Log_{1000}DKDB$
2000	70	3.182	0.002	0.168	2011	3 338	303.455	0.448	0.827
2001	166	7.905	0.006	0.299	2012	3 029	302.900	0.526	0.827
2002	673	33.650	0.023	0.509	2013	3 244	360.444	0.608	0.852
2003	1 375	72.368	0.058	0.620	2014	3 100	387.500	0.688	0.863
2004	539	29.944	0.072	0.492	2015	3 011	430.143	0.764	0.878
2005	1 034	60.824	0.098	0.595	2016	2 453	408.833	0.827	0.871
2006	1 367	85.438	0.133	0.644	2017	2 452	490.400	0.890	0.897
2007	1 674	111.600	0.176	0.683	2018	1 913	478.250	0.938	0.893
2008	2 056	146.857	0.228	0.722	2019	1 486	495.333	0.976	0.898
2009	2 242	172.462	0.286	0.746	2020	850	425.000	0.998	0.876
2010	3 039	253.250	0.363	0.801	2021	78	78.000	1.000	0.631

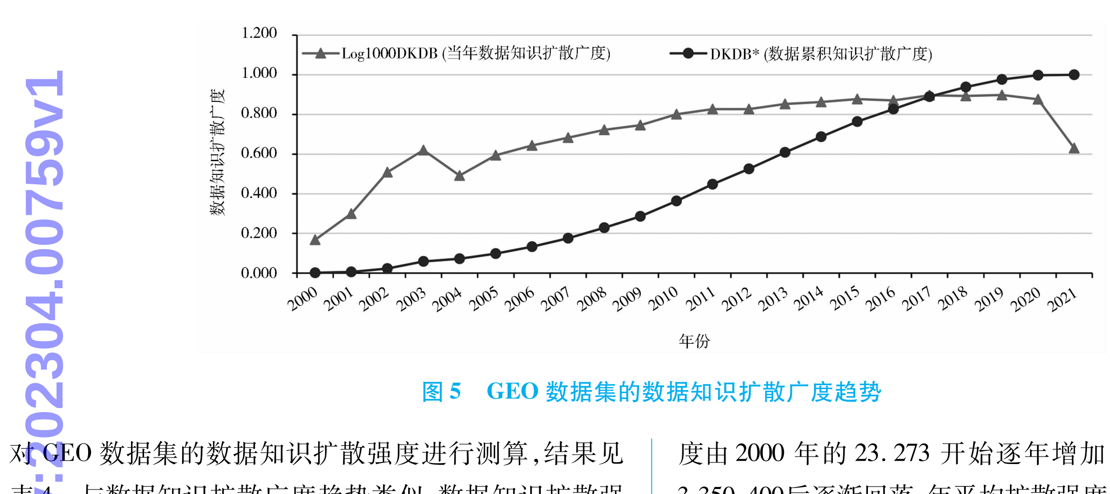


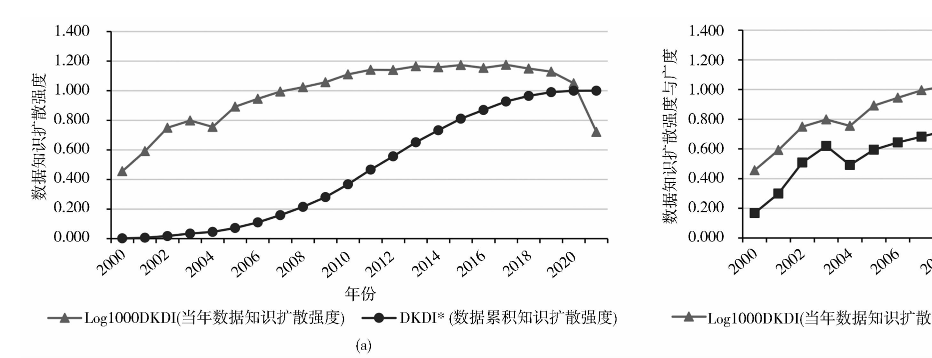
图 5　GEO 数据集的数据知识扩散广度趋势

对 GEO 数据集的数据知识扩散强度进行测算,结果见表4。与数据知识扩散广度趋势类似,数据知识扩散强度由 2000 年的 23.273 开始逐年增加,到 2017 年达到 3 350.400后逐渐回落,年平均扩散强度为 1 607.756。

表 4　GEO 数据集的数据知识扩散强度

年度	N_j	DKDI	DKDI*	$Log_{1000}DKDI$	年度	N_j	DKDI	DKDI*	$Log_{1000}DKDI$
2000	512	23.273	0.002	0.456	2011	29 050	2 640.909	0.466	1.141
2001	1251	59.571	0.006	0.592	2012	26 345	2 634.500	0.556	1.140
2002	3 548	177.400	0.018	0.750	2013	28 022	3 113.556	0.651	1.164
2003	4 721	248.474	0.034	0.798	2014	23 838	2 979.750	0.732	1.158
2004	3 314	184.111	0.045	0.755	2015	23 209	3 315.571	0.811	1.174
2005	8 088	475.765	0.073	0.892	2016	17 418	2 903.000	0.870	1.154
2006	10 952	684.500	0.110	0.945	2017	16 752	3 350.400	0.927	1.175
2007	14 452	963.467	0.159	0.995	2018	11 252	2 813.000	0.965	1.150
2008	16 423	1 173.071	0.215	1.023	2019	7 322	2 440.667	0.990	1.129
2009	19 346	1 488.154	0.280	1.058	2020	2 822	1 411.000	0.999	1.050
2010	25 734	2 144.500	0.368	1.110	2021	146	146.000	1.000	0.721

本文以计算出的 *DKDI* 取 1 000 为底的对数作为当年数据知识扩散指标,对当年数据知识扩散与数据累积知识扩散指标的结果进行比较和分析,见图 6(a)。可以看出,2000－2017 年间的数据知识扩散强度波动增长,而由于数据集的滞后性,其后逐渐回落。计算 *DKDI* 和 *DKDB* 的相关系数为 0.998,二者呈现出高度相关性。*DKDI* 和 *DKDB* 的年度增长趋势对比见图 6(b),由图 6(b)可以看出二者几乎保持着一致的变化趋势,随着序列比对、基因识别等生物医学领域的研究不断深入,论文对科学数据集的使用强度也随之加大,科研人员对数据的依赖性日益增加。



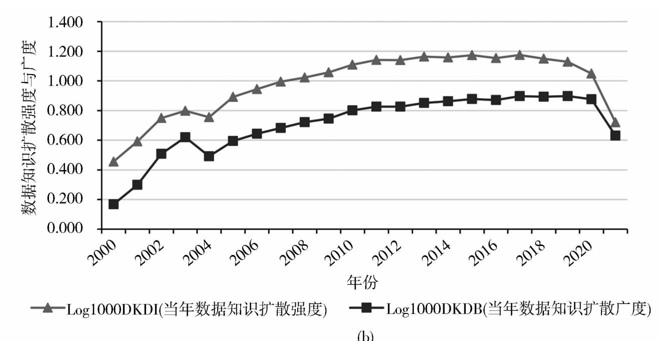


图 6 GEO 数据集的数据知识扩散强度趋势及与扩散广度趋势对比

3.4 数据知识扩散速度

数据知识扩散速度可以反映出研究人员对数据的关注和利用效率，对于扩散目标来说，消耗的时间越短，扩散速度越快。而更快的传播和利用速度可以有效减少知识老化导致的学术价值损失，有效降低知识创新成本，从而加快科技发展速度。数据知识扩散速度最低值为 2005 年的 0.006，最高为 2021 年的 0.515，年平均扩散速度为 0.040，如图 7 所示：



图 7 GEO 数据集的数据知识扩散速度趋势

由图 7 可知，数据在初期得到了广泛的关注和利用，而到了 2003 年以后，由于基因测序技术的进步以及成本的不断降低，科研人员开始通过自身实验来获取数据，数据知识扩散速度开始放缓。直到 2014 年后，随着数据库内容的不断完善，科研人员又逐步开始通过重用他人数据的形式进行研究，重用数据可以大大减少科研成本，加快科研进度。

3.5 数据知识扩散延时

数据知识扩散延时从数据初次扩散耗时的角度，即从重用数据到文献发表的周期，对数据知识扩散的速度进行揭示。经过计算发现，GEO 数据集的最大扩散延时为 20 年，最小扩散延时为 0 年，平均扩散延时约为 3.8 年。如果以数据集公开年份与数据集最初提交年份的差值，作为数据贡献者从数据处理到论文发表的延时，则可以得到原始数据处理到文献发表的平均周期为 4.3 年左右，重用数据可以将科研效率平均提升 13.16%。二者的分布如图 8 所示：

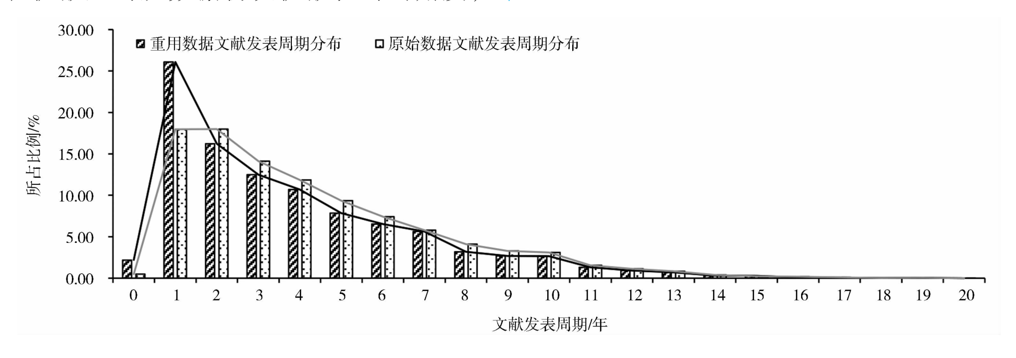


图 8 重用数据文献发表周期与原始数据文献发表周期分布

由图 8 可以看出，生物医学领域的数据知识扩散延时为 1 的占比最高，约为 26.08%，说明有超过 1/4 的数据在公开一年后就被其他发表的文献所重用。而原始数据文献发表周期为 2 的占比最高，约为 17.99%，说明从原始数据处理到文献发表的周期数量最多的是 2 年。从数据知识扩散延时平均数 4 以内的总体情况来看，重用数据的文献发表周期在 0-4 年的占比为 57%，高于原始数据文献发表周期在 0-4 年的占比 50.58%，再次印证了重用数据可以一定程度上缩短文献发表周期，加快知识扩散速度。

4 讨论

本研究以 GEO 数据库中的科学数据集和 PMC 数据库的全文数据为研究对象，分析了生物医学领域的基因表达数据集共享及重用的情况，并针对科学数据集的特点提出了数据知识扩散测度指标，最后利用共享和重用数据信息研究了科学数据集的知识扩散特征。研究结果不仅探析到科学数据集在学科研究过程中产生的实际价值，也发现了数据集的重用行为可以提高科研效率、加速科研进程，通过定量分析证实了科学数据集在生物医学领域研究中的重要地位和作用，得到的具体结论如下：

（1）GEO 数据库中被重用较多的数据集集中在 2008-2017 年期间，这与近 10 年来数据科学由知识范式到数据范式的数据科学的发展转变历程基本吻合。此外，由于数据重用存在着一定的滞后性，被重用的数据集数量在时间上呈持续增长又逐步下降的趋势，这与数据知识扩散延时约为 4 年的结论相符合，体现出较明显的规律性。

（2）2001-2021 年 GEO 数据集的当年数据知识扩散广度和当年数据知识扩散强度都保持着波动增长，又在近几年有所下降的态势。从累积数据知识扩散广度和累积数据知识扩散强度来看，二者都呈现出 S 曲线形态，符合科学数据集公开和重用具有一定时间滞后性的特点。总体来看，无论从数据集在短期还是长期产生的知识价值方面，数据集对于生物医学领域科研产生的影响力都日渐增长，科研人员在研究中使用的数据集数量也在不断增多，体现了生物医学领域由数据驱动科研的学科特点。

（3）从 GEO 数据集的数据知识扩散速度角度来看，2001-2021 年 GEO 数据集的数据知识扩散经历了波动、平稳再到提速的 3 个主要阶段。初期由于数据共享这一新的科研方式出现并且数量较少，数据集快速得到了大量关注和使用。而到了 2003 年以后，随着测序成本的不断降低，科研人员更倾向于通过自身实验产生并共享数据，数据知识扩散速度开始且呈平稳发展态势。随着数据库的不断完善以及数据集的不断积累，科研人员又重新开始关注并重用数据集，尤其是 2014 年后，数据知识扩散速度开始呈现指数级增长态势。

（4）由于我国在科学数据共享方面起步较晚，初期共享数据集较多的机构大多集中在与国外科研合作较多的香港等地区。随着国家和科研机构对科研数据管理与共享的重视程度不断增强，以及《科学数据管理办法》《中国科学院科学数据管理与开放共享办法》等政策的不断出台，我国在科学数据领域的地位不断提高，这些工作为我国建设成为数据强国奠定了坚实的基础和保障。

5 总结

本研究将以生物医学领域的基因表达数据集作为研究对象，将科学数据集这一新的知识实体纳入知识扩散的研究范畴，提出适用于科学数据集知识扩散研究的测度指标，从而揭示科学数据集在参与科研过程中的特点和规律。同前人的研究相比，本研究拓展了知识扩散的研究理论和方法，可以为科学数据管理和服务工作提供参考依据。首先，要进一步推进科学数据管理和共享政策的不断完善，提高我国科研工作者的数据共享和重用意识。随着我国科学数据管理政策的陆续出台和科研机构对于科学数据重视程度的增强，我国在国际科学数据舞台上也逐渐成为主角，要继续保持这种良好发展态势，不断夯实我国作为科学数据强国的主导地位；其次，我国科学数据库建设要更加强调专业性、及时性和开放性，专业性的数据库具备更强的吸引力，不但要将宝贵的科学数据留在祖国大地上，更要吸引全球的数据流入和汇聚在我国的科学数据库中。这就要求科学数据库建设要聘请专业运维团队及同行评议专家进行及时更新维护，并且通过多渠道资金优化配置保证数据的免费和开放共享；最后，高校和图书馆要加强科学数据人才培养，专业型人才渗透到科学数据集产生、共享、重用等各个环节，加快科学数据集知识扩散过程，让其在科研过程中发挥更大

的作用,满足飞速发展的科学数据管理和服务需求。

当然,本文研究也存在着一些不足,亟待进一步研究。首先,仅以生物医学领域基因表达数据集和 GEO 数据库为例,学科和样本数据都还有待进一步丰富和加强;其次,未深入分析发布者、发布机构、原文影响因子、国家地区等因素与数据集影响力和扩散特征之间的关联性;最后,还可以从更多样化的视角探索科学数据集的知识扩散特征,如基于网络结构的数据集扩散特征、基于合作关系的数据集扩散特征等。

参考文献:

[1] CHEN C M, HICKS D. Tracing knowledge diffusion[J]. Scientometrics, 2004, 59(2): 199-211.

[2] LEARNED W S. The American public library and the diffusion of knowledge[J]. Journal of the American Medical Association, 1924, 83(20): 1611-1611.

[3] 黄鲁成,刘玉敏,吴菲菲,等. 基于专利全引用信息的技术知识扩散特征研究——以石墨烯技术为例[J]. 科学学与科学技术管理, 2017, 38(4): 149-161.

[4] 赵蓉英,魏绪秋. 引证视角下的作者知识扩散规律探析[J]. 情报理论与实践, 2016, 39(8): 12-17.

[5] 岳增慧,许海云. 学科引证网络知识扩散特征研究[J]. 情报学报, 2019, 38(1): 1-12.

[6] 王静静,叶鹰. 国际数字人文研究中的跨学科知识扩散探析[J]. 大学图书馆学报, 2021, 39(2): 45-51, 61.

[7] LIU Y X, ROUSSEAU R. Knowledge diffusion through publications and citations: a case study using esi-fields as unit of diffusion[J]. Journal of the American Society for Information Science and Technology, 2010, 61(2): 340-351.

[8] 俞立平,万晓云,项益鸣,等. 一个评价学术期刊知识扩散深度的新指标——cjh 指数[J]. 情报杂志, 2019, 38(8): 145-150.

[9] NAKAMURA H, SUZUKI S, HIRONORI T, et al. Citation lag analysis in supply chain research[J]. Scientometrics, 2011, 87(2): 221-232.

[10] 宋歌. 学术创新的扩散过程研究[J]. 中国图书馆学报, 2015, 41(1): 62-75.

[11] KISS I Z, BROOM M, CRAZE P, et al. Can epidemic models describe the diffusion of topics across disciplines? [J]. Journal of informetrics, 2010, 4(1): 74-82.

[12] GAO X, GUAN J C. Network model of knowledge diffusion[J]. Scientometrics, 2012, 90(3): 749-762.

[13] 魏绪秋,郭凤娇,于森. 微观视域下的图书知识扩散特征探析[J]. 情报科学, 2021, 39(3): 37-43.

[14] 于晓彤,潘雪莲,华薇娜. 知识图谱研究中的软件引用和扩散分析[J]. 情报资料工作, 2019, 40(2): 19-29.

[15] 张玲玲,张宇娥,杜丽. 国家社科基金项目成果视角下图情领域知识扩散研究[J]. 图书馆工作与研究, 2017(10): 60-66.

[16] PARK H, YOU S, WOLFRAM D. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields[J]. Journal of the Association for Information Science and Technology, 2018, 69(11): 1346-1354.

[17] 孟祥保,钱鹏. 数据生命周期视角下人文社会科学数据特征研究[J]. 图书情报知识, 2017(1): 76-88.

[18] 丁文姚,李健,韩毅. 我国图书情报领域期刊论文的科学数据引用特征研究[J]. 图书情报工作, 2019, 63(22): 118-128.

[19] 刘亚男,刘江荣,肖明,等. 基金项目论文中的科研数据引用行为研究[J]. 图书馆论坛, 2019, 39(7): 75-83.

[20] ZHAO M N, YAN E J, LI K. Data set mentions and citations: a content analysis of full-text publications[J]. Journal of the Association for Information Science and Technology, 2018, 69(1): 32-46.

[21] 闫小妮,田国祥,郭晓娟,等. Geo 数据库架构、申请及数据提取方法与流程[J]. 中国循证心血管医学杂志, 2019, 11(2): 134-137.

[22] 王雪,杨波. 科学数据重复使用的学科差异性研究[J]. 情报杂志, 2021, 40(7): 122-126+156.

[23] 阮继,王玥,刘谦,等. Pubmed central 引用数据在中文科技期刊平台展示的实现[J]. 科技与出版, 2020(3): 125-128.

[24] ROWLANDS I. Journal diffusion factors: a new approach to measuring research influence[J]. Aslib proceedings, 2002, 54(2): 77-84.

[25] FRANDSEN T F, ROUSSEAU R, ROWLANDS I. Diffusion factors[J]. Journal of documentation, 2006, 62(1): 58-72.

[26] 邱均平,瞿辉,罗力. 基于期刊引证关系的学科知识扩散计量研究——以我国"图书馆、情报、档案学"为例[J]. 情报科学, 2012, 30(4): 481-485, 491.

[27] 李江. 基于引文的知识扩散研究评述[J]. 情报资料工作, 2013(4): 36-40.

[28] 汤易兵,黄祖庆,张宝友. 基于引文网络的知识扩散和整合研究——以供应链研究为例[J]. 情报杂志, 2012, 31(1): 119-122.

[29] OSIER M V, ZHAO H Y, CHEUNG K H. Handling multiple testing while interpreting microarrays with the gene ontology database[J]. Bmc bioinformatics, 2004, 5.

[30] GEO. Development stages of lentinula edodes[EB/OL]. [2021-11-12]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2167.

作者贡献说明:

杨宁:资料搜集,实验验证及初稿撰写;

张志强:指导论文修改,凝练论文研究点。

杨宁，张志强.科学数据集知识扩散特征探析——以基因表达数据集为例[J].图书情报工作,2022,66(12):82-91.

Research on the Characteristics of Knowledge Diffusion in Scientific Datasets ——Taking the Gene Expression Dataset as an Example

Yang Ning[1,2] Zhang Zhiqiang[1,2]

[1] Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041

[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/Significance] By studying the characteristics and laws of knowledge diffusion of scientific datasets, this paper explores the practical role of scientific datasets in the development of discipline fields, so as to provide references for scientific and technological evaluation and management policy-making of scientific datasets. [Method/Process] Taking the datasets of GEO database and the full-text data of reused dataset in PubMed Central Database as the analysis objects, this paper analyzed the knowledge diffusion characteristics of scientific datasets by using content analysis method combined with knowledge diffusion indicators such as diffusion breadth, diffusion intensity and diffusion speed. [Result/Conclusion] The results show that the breadth and intensity of knowledge diffusion of scientific datasets are increasing day by day. Reusing data can accelerate the speed of knowledge diffusion, and China's position in the field of global scientific data is improving.

Keywords: scientific dataset knowledge diffusion analysis of characteristics measure index